# SliceMatch: Geometry-guided Aggregation for Cross-View Pose Estimation Supplementary Material

Ted Lentsch* Zimin Xia* Holger Caesar Julian F. P. Kooij
Intelligent Vehicles Group, Delft University of Technology, The Netherlands
{T.deVriesLentsch,Z.Xia,H.Caesar,J.F.P.Kooij}@tudelft.nl

## Overview

In this supplementary material, we provide the following items for a better understanding of the main paper:

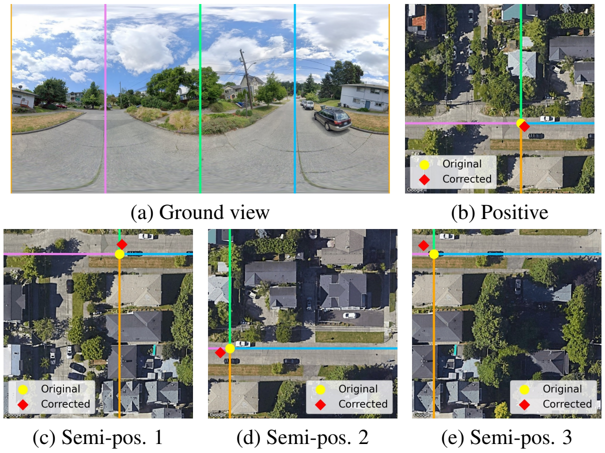
A. Ground Truth Labels of VIGOR Dataset (*Section 4.1*)
B. Tuning Number of Slices (*Section 4.5*)
C. Inference on Images with a Limited HFoV (*Section 4.6*)
D. Details on Runtime and Memory Usage (*Section 4.8*)
E. Visualization: SliceMatch Predictions (*Section 4.6*)

## A. Ground Truth Labels of VIGOR Dataset

We have visually inspected the image pairs of the VIGOR dataset [4] and noticed a location inconsistency between image pairs that share the ground image. Figure 1 shows an image pair formed by a ground image and a positive or semi-positive aerial image from Seattle with the original and corrected ground truth camera locations indicated. Depending on the aerial image, the original ground truth location (yellow dot) is in different locations. However, this should be the same visual location (red diamond) for all aerial images corresponding to this specific ground image.

The authors of the VIGOR dataset [4] have used a ground resolution equal to 0.114m/pixel for all 4 cities of the dataset to convert the latitude and longitude of a ground image to its location in aerial images. We have measured the ground resolution ourselves. Pixel-level correspondences between aerial images that have a visual overlap can be calculated using cross-correlation. Then we can overlay these aerial images. Figure 2 shows this for two aerial images. The distance in pixels between the two image centers can be measured (see Figure 2c). In addition, the longitude and latitude of the image center of each aerial image are known, allowing the distance to be determined in meters as well. The ground resolution of an aerial-aerial image combination can be calculated using,

$$\textit{ground resolution} = \frac{\textit{distance in meters}}{\textit{distance in pixels}}. \tag{1}$$



(a) Ground view (b) Positive

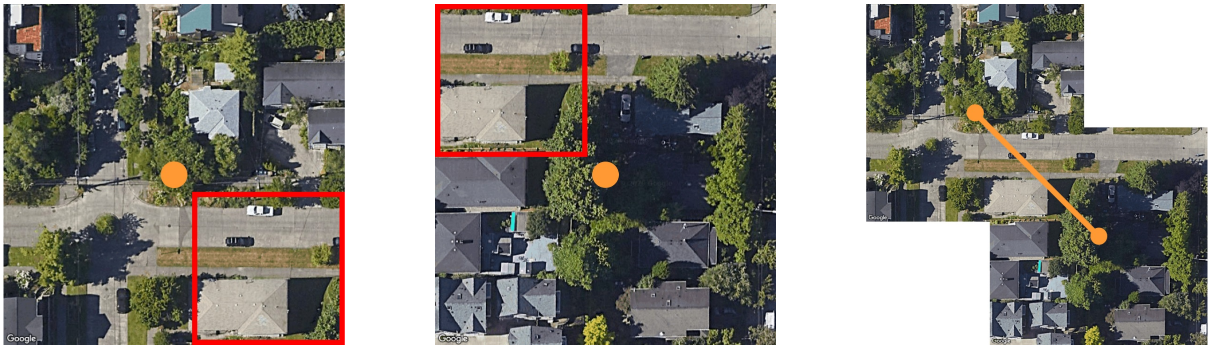
(c) Semi-pos. 1 (d) Semi-pos. 2 (e) Semi-pos. 3

Figure 1. **A ground image together with the four matching aerial images from VIGOR [4].** The original and the corrected locations are indicated by the yellow dot and red diamond, respectively. South, West, North, and East are the orange, pink, green, and blue lines, respectively. *Semi-pos.* means *Semi-positive*.

We have calculated a new ground resolution for each city by averaging the ground resolutions of a city's aerial-aerial image combinations. Table 1 shows the original and our measured ground resolutions. It turns out that the ground resolutions differ per city. The measured value for the ground resolution for New York is almost equal to the ground resolution that the VIGOR authors have used. However, the measured resolutions for the other cities differ significantly.

The measured ground resolutions have been used to determine corrected ground truth location labels. Table 2 shows statistics on the absolute error in meters between the original and corrected locations. The positive image pairs of the dataset were used to determine the statistics since only those image pairs were used for the experiments (see Section 4.1). For Seattle, the difference between the original and measured ground resolution is the largest and this results in errors of more than 3 meters (see Table 1). The other 3 cities have smaller mean and median errors than Seattle.

[*] indicates equal contribution.



(a) Aerial image 1

(b) Aerial image 2

(c) The overlaid aerial image 1 and 2

Figure 2. **Two aerial images from the VIGOR dataset [4] that have a visual overlap.** Each image center is indicated by an orange dot and the connection between the two dots shows the distance in pixels. (a) Aerial image 1 with the visual overlap indicated by the red contour. (b) Aerial image 2 with the visual overlap indicated by the red contour. (c) The overlaid aerial image 1 and 2. We use cross-correlation to find the amount of overlapping (in pixels) between aerial image 1 and 2.

In our localization only and pose estimation experiments for the VIGOR dataset, we resize the aerial image to $512 \times 512$ pixels. As a result, the ground resolution of the resized aerial images can be obtained by multiplying the measured ground resolutions from Table 1 by 1.25 $(= 640/512)$.

| City | Original | Measured |
|---|---|---|
| Chicago | 0.114 | 0.111 |
| New York | 0.114 | 0.113 |
| San Francisco | 0.114 | 0.118 |
| Seattle | 0.114 | 0.101 |

Table 1. **The original and our measured ground resolution for the 4 cities from VIGOR [4].** The ground resolutions correspond to aerial images with a size of $640 \times 640$ pixels, and the unit of the ground resolution is m/pixel.

| City | Min. | Mean | Median | Max. |
|---|---|---|---|---|
| Chicago | 0.00 | 0.43 | 0.45 | 0.80 |
| New York | 0.00 | 0.25 | 0.25 | 0.47 |
| San Francisco | 0.00 | 0.46 | 0.49 | 0.95 |
| Seattle | 0.00 | 1.72 | 1.79 | 3.14 |

Table 2. **Statistics on the absolute error in meters of the labels for the 4 cities from VIGOR [4].** The absolute error is defined as the distance between the original and the corrected locations. *Min.* and *Max.* indicate *Minimum* and *Maximum*, respectively.

## B. Tuning Number of Slices

To supplement our ablation study on the number of slices (see Section 4.5), we visualize the predictions from SliceMatch models with different numbers of slices on VIGOR same-area, see Figure 3. Using larger $N$ (more slices) maintains more of the relative orientation between the visible components in the scene. Note that lowering $N$ makes the descriptors less orientation aware, which we observe lowers performance. Generally, we observe that increasing to $N = 16$ makes the descriptors more discriminative, resulting in less uncertainty about the true location and orientation. However, our ablation study in the main paper Table 1 demonstrates a trade-off: if $N$ becomes too large, the descriptor becomes too sensitive to pose differences between the best candidate pose and true pose.

Similarly, we also conducted an ablation study for the number of slices on the KITTI dataset [1, 2], see Table 3. For this study, we used the Same-Area setting of KITTI and the $20^\circ$ orientation prior. Similar to the VIGOR dataset, the highest performance is achieved with 16 slices for the KITTI dataset as well. For 8 and 32 slices, the performance is slightly worse.

|  | Cross-View | ↓ Location (m) |  | ↓ Orientation (°) |  |
|---|---|---|---|---|---|
| N | Attention | Mean | Median | Mean | Median |
| 8 | ✓ | 8.74 | 5.11 | 4.54 | 4.01 |
| 16 | ✓ | **7.96** | **4.39** | **4.12** | **3.65** |
| 32 | ✓ | 8.03 | 4.72 | 4.34 | 3.65 |

Table 3. **Location and orientation error for different slice number $N$ values for the Same-Area setting on the KITTI dataset [1, 2].** Best performance in **bold**.

## C. Inference on Images with a Limited HFoV

Additionally, we conducted experiments that vary the HFoV of test images in the VIGOR dataset (same-area), see Figure 4 (median errors). As expected, SliceMatch's performance degrades when the HFoV of the ground-level query image reduces, as it contains less information. Training on

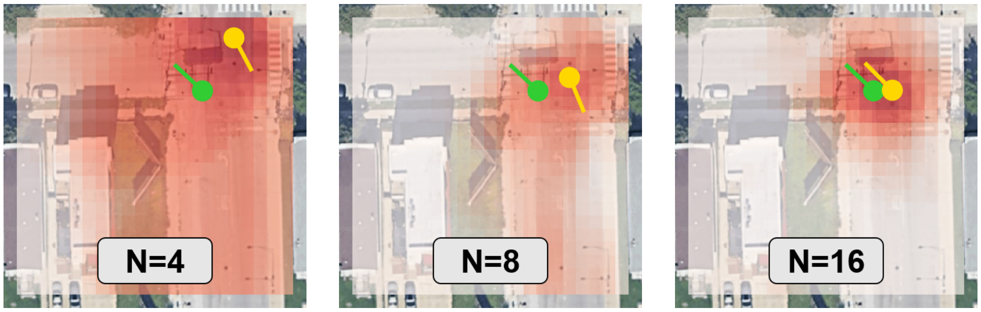
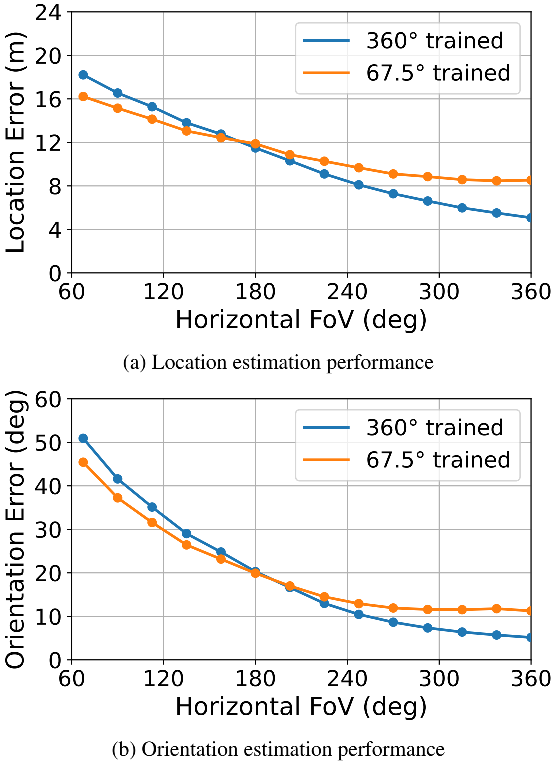


Figure 3. **SliceMatch models with different slice number $N$ values.** The ground truth camera pose and our estimated camera pose are in green and yellow, respectively. Red shading indicates the highest similarity score between the ground descriptor and the aerial descriptors among all orientations at that location.

ground images with a small HFoV, e.g. $\sim 67.5°$, recuperates some performance when testing on small HFoVs.



(a) Location estimation performance

(b) Orientation estimation performance

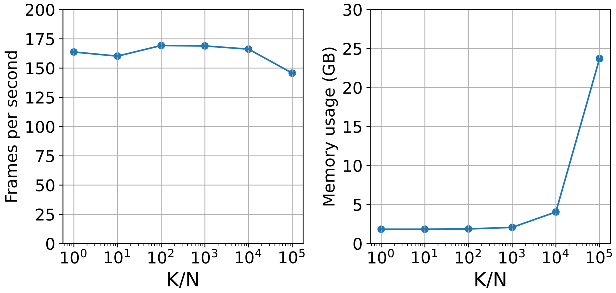
Figure 4. **Median location and orientation estimation errors on VIGOR [4] for limited HFoV.** (a) Location estimation performance. (b) Orientation estimation performance.

## D. Details on Runtime and Memory Usage

Here, we provide a detailed analysis of the runtime and memory usage of SliceMatch (see Section 4.8).

Both pre-computation of slice masks and parallelization of pose descriptor aggregation contribute to our efficiency. For a single input pair, we process candidate poses in parallel by performing a single (large) matrix multiplication (the matrix has the pre-computed masks as rows). Note that the candidate poses and thus the masks are the same for all test images. Since we never need to recompute the masks, precomputation is excluded from the reported inference time. On a single input pair in the VIGOR dataset, our *feature extraction* / *descriptors aggregation* / *pose comparison* takes 3.4ms / 1.5ms / 1.1ms, respectively. For the KITTI dataset, this is 3.6ms / 1.6ms / 1.2ms, respectively.

Importantly, the runtime of SliceMatch remains nearly constant as the number of used candidate poses $K$ increases, while memory scales linearly, see Figure 5 where we test for $N = 16$ and $K/N \in \{1, 10, 100, 1k, 10k, 100k\}$. Our main experiments used $K/N = 28.2k/16 = 1.76k$ for VIGOR, and $K/N = 14.4k/16 = 0.9k$ for KITTI. In practice, memory will thus be the limiting factor for determining the number of poses that can be used.



(a) Frames per second

(b) Memory usage

Figure 5. **SliceMatch frames per second and memory usage for a varying number of poses.** Note that the x-axis uses log-scale.

## E. Visualization: SliceMatch Predictions

Here we provide extra qualitative results for our experiments in the main paper (see Section 4.6).

Figure 6 shows successful predictions of SliceMatch for the VIGOR [4] and KITTI [1, 2] datasets. In Figure 6a, Figure 6b, Figure 6c, and Figure 6h, it can be seen that the predicted similarity map is aligned with the road and that the predicted orientation is in line with the orientation of the ground image. In contrast, in Figure 6a, Figure 6b, Figure 6g, and Figure 6h, MCC [3] predicts a location on the road, but the orientation sometimes differs 180 degrees from that of the ground image. In Figure 6d and Figure 6i, LM [2] converges to a location on the roof of a building for some image pairs.

Figure 7 shows some failure cases of SliceMatch. SliceMatch can predict a multi-modal similarity map. In Figure 7a, Figure 7b, Figure 7c and Figure 7d, it can be seen that SliceMatch predicts two peaks, but the wrong peak is used as the prediction. The ground image of Figure 7e contains few discriminating objects and this can be observed in the predicted similarity map. SliceMatch predicts uncertainty aligned with the road. The lateral error is small, however, the longitudinal error is large.

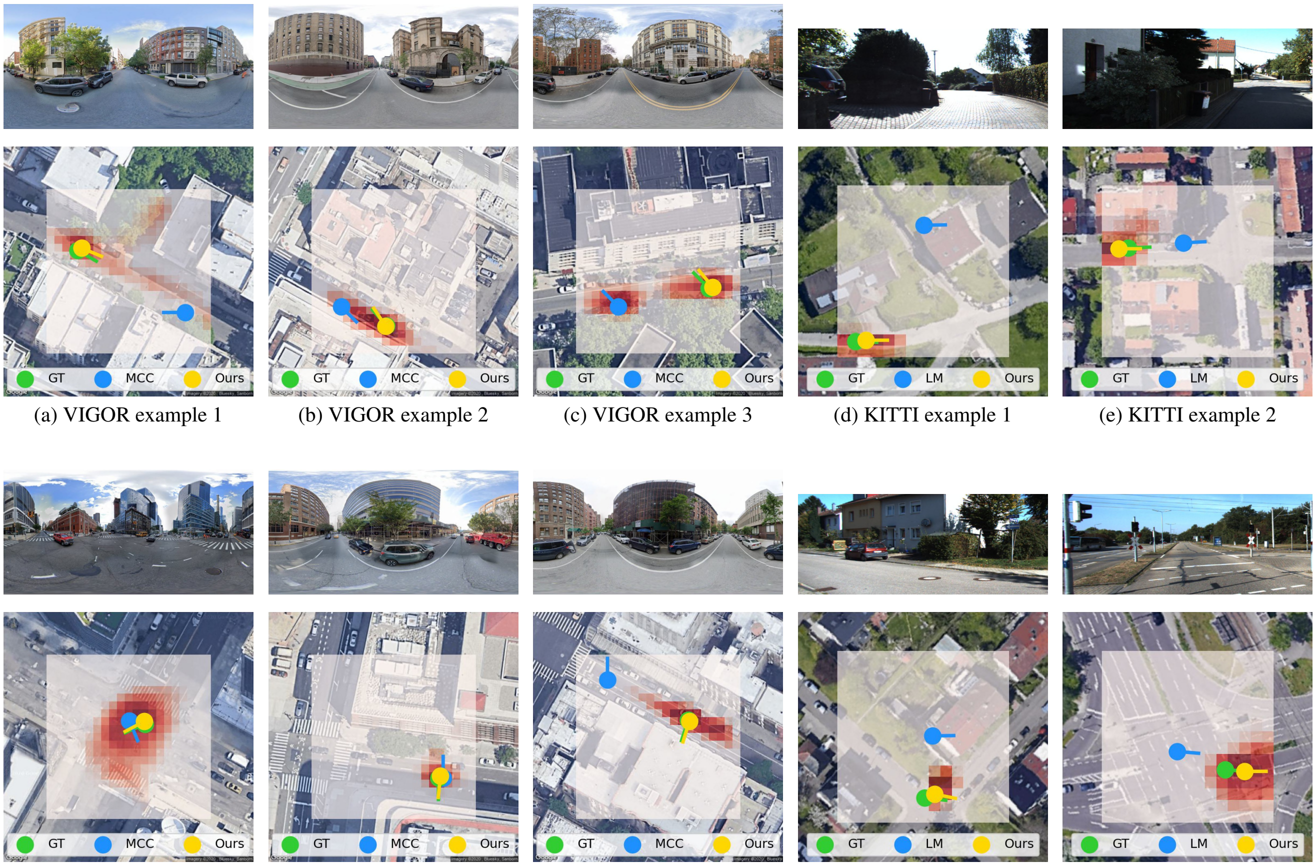
## References

[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11):1231–1237, 2013. 2, 4, 5

[2] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *CVPR*, pages 17010–17020, 2022. 2, 4, 5

[3] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *ECCV*, pages 90–106, 2022. 4

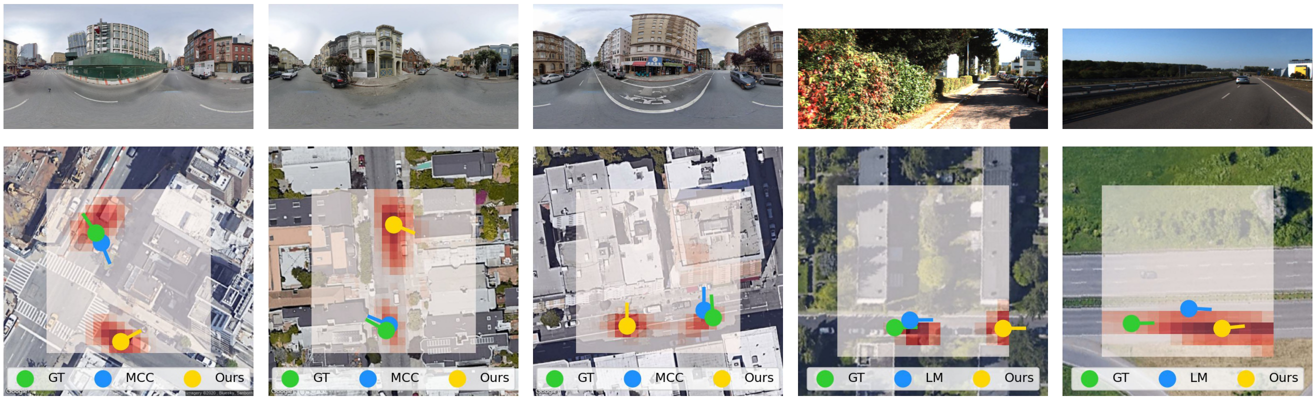
[4] Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *CVPR*, pages 3640–3649, 2021. 1, 2, 3, 4, 5



(a) VIGOR example 1 (b) VIGOR example 2 (c) VIGOR example 3 (d) KITTI example 1 (e) KITTI example 2

(f) VIGOR example 4 (g) VIGOR example 5 (h) VIGOR example 6 (i) KITTI example 3 (j) KITTI example 4

Figure 6. **Qualitative evaluation of SliceMatch on VIGOR [4] and KITTI [1, 2]: successful cases.** Top row: input ground image. Bottom row: GT and pose estimation results overlayed on input aerial image. Red shading indicates the highest similarity score between the ground descriptor and the aerial descriptors among all orientations at that location.



(a) VIGOR example 1 (b) VIGOR example 2 (c) VIGOR example 3 (d) KITTI example 1 (e) KITTI example 2

Figure 7. **Qualitative evaluation of SliceMatch on VIGOR [4] and KITTI [1, 2]: failure cases.** Top row: input ground image. Bottom row: GT and pose estimation results overlayed on input aerial image. Red shading indicates highest similarity score between the ground descriptor and the aerial descriptors among all orientations at that location.